ABSTRAK

KLASIFIKASI SITUS HIDROKSILASI PROTEIN PADA PROLIN DAN LISIN MENGGUNAKAN EXTREME GRADIENT BOOSTING (XGBOOST)

Oleh

Intania Rahmadhilla

Modifikasi pasca-translasi (PTM) memiliki peran penting dalam berbagai sel dan proses biologis. Sebagian besar protein menjalankan fungsi biologisnya setelah melalui modifikasi pasca-translasi. Hidroksilasi protein merupakan salah satu jenis PTM yang terjadi pada prolin (P) dan lisin (K). Diketahui bahwa sekuens protein mengandung banyak residu P dan K yang tidak dikarakterisasi, pertanyaannya mana yang bisa dihidroksilasi dan mana yang tidak bisa. Untuk mengatasi hal tersebut, metode komputasi sangat disarankan untuk mengidentifikasi situs hidroksilasi karena tidak memakan waktu dan biaya yang mahal. Distribusi kelas yang tidak seimbang menjadi tantangan berat bagi algoritma pembelajaran tradisional untuk meningkatkan hasil akurasi keseluruhan, sehingga penelitian ini menggunakan pendekatan machine learning dengan metode Extreme Gradient Boosting (XGBoost) dan implementasi algoritma SMOTE untuk mengatasi kelas data tidak seimbang. Fitur ekstraksi pada penelitian ini terdiri atas tiga macam fitur, yaitu Pseudo Amino Acid Composition (PseAAC), CTD, dan AAindex. Berdasarkan dataset uji, hasil akurasi mencapai 97.9% dan 99.1% untuk prediksi situs hidroksilasi prolin dan hidroksilasi lisin. Sementara itu, diperoleh sensitivitas sebesar 93.6% dan 95.4%, f-1 score sebesar 92.9% dan 97.6%, serta MCC 95.3% dan 97.1% untuk residu P dan K.

Kata Kunci : PTM, hidroksilasi, data tidak seimbang, fitur ekstraksi, xgboost, SMOTE

ABSTRACT

CLASSIFICATION OF PROTEIN HYDROXYLATION SITES ON PROLINE AND LYSINE USING EXTREME GRADIENT BOOSTING (XGBOOST)

By

Intania Rahmadhilla

Post-translational modification (PTMs) has an important role in various cells and biological processes. Most proteins perform their biological functions after going through post-translational modifications. Protein hydroxylation is a type of PTM that occurs in proline (P) and lysine (K). It is known that protein sequences contain many uncharacterized residues of P and K, the question is which ones can be hydroxylated and which ones can't. To overcome this, computational methods are highly recommended to identify hydroxylation sites because they are not time-consuming and costly. Imbalanced class distribution present a major challenge for traditional learning algorithms to improve overall accuracy results. Therefore in this study, we use an Extreme Gradient Boosting (XGBoost) machine learning approach and the implementation of SMOTE algorithm to overcome imbalanced data classes. The extraction features in this study consist of three kinds of features, namely Pseudo Amino Acid Composition (PseAAC), CTD, and AAindex. Based on the test data set, the accuracy results reached 97.9% and 99.1% for the prediction of proline hydroxylation sites and lysine hydroxylation. Meanwhile, sensitivity was obtained by 93.6% and 95.4%, f-1 scores of 92.9% and 97.6%, and MCC of 95.3% and 97.1% for residues of P and K.

Keywords : PTMs, hydroxylation, imbalanced data, feature extraction, xgboost, SMOTE